Rajesh Singh
Department of Statistics, Banaras Hindu University(U.P.), India

Pankaj Chauhan, Nirmala Sawan
School of Statistics, DAVV, Indore (M.P.), India

Florentin Smarandache
Department of Mathematics, University of New Mexico, Gallup, USA

Improvement in Estimating Population Mean using Two Auxiliary Variables in Two-Phase Sampling

Published in:
Rajesh Singh, Jayant Singh, F. Smarandache (Editors)
STUDIES IN STATISTICAL INference, SAMPLING TECHNIQUES AND DEMOGRAPHY
InfoLearnQuest, Ann Arbor, USA, 2009
pp. 26 - 35
Abstract

This study proposes improved chain-ratio type estimator for estimating population mean using some known values of population parameter(s) of the second auxiliary character. The proposed estimators have been compared with two-phase ratio estimator and some other chain ratio type estimators. The performances of the proposed estimators have been supported with a numerical illustration.

Key words: Auxiliary variables, chain ratio-type estimator, bias, mean squared error.

1. Introduction

The ratio method of estimation is generally used when the study variable Y is positively correlated with an auxiliary variable X whose population mean is known in advance. In the absence of the knowledge on the population mean of the auxiliary character we go for two-phase (double) sampling. The two-phase sampling happens to be a powerful and cost effective (economical) procedure for finding the reliable estimate in first phase sample for the unknown parameters of the auxiliary variable x and hence has eminent role to play in survey sampling, for instance, see Hidiroglou and Sarndal (1998).
Consider a finite population \( U = (U_1, U_2, \ldots, U_N) \). Let \( y \) and \( x \) be the study and auxiliary variable, taking values \( y_i \) and \( x_i \) respectively for the \( i \)th unit \( U_i \).

Allowing SRSWOR (Simple Random Sampling without Replacement) design in each phase, the two-phase sampling scheme is as follows:

(i) the first phase sample \( s'_n \) (\( s'_n \subset U \)) of a fixed size \( n' \) is drawn to measure only \( x \) in order to formulate a good estimate of a population mean \( \bar{X} \),

(ii) Given \( s'_n \), the second phase sample \( s_n \) (\( s_n \subset s'_n \)) of a fixed size \( n \) is drawn to measure \( y \) only.

Let \( \bar{x} = \frac{1}{n} \sum_{i \in s_n} x_i \), \( \bar{y} = \frac{1}{n} \sum_{i \in s_n} y_i \) and \( \bar{x}' = \frac{1}{n'} \sum_{i \in s_n} x_i \).

The classical ratio estimator for \( Y \) is defined as

\[
\bar{y}_r = \frac{\bar{y}}{\bar{x}} \tag{1.1}
\]

If \( \bar{X} \) is not known, we estimate \( \bar{Y} \) by two-phase ratio estimator

\[
\bar{y}_{tr} = \frac{\bar{y}}{\bar{x}'} \tag{1.2}
\]

Some times even if \( \bar{X} \) is not known, information on a cheaply ascertainable variable \( z \), closely related to \( x \) but compared to \( x \) remotely related to \( y \), is available on all units of the population. For instance, while estimating the total yield of wheat in a village, the yield and area under the crop are likely to be unknown, but the total area of each farm may be known from village records or may be obtained at a low cost. Then \( y \), \( x \) and \( z \) are respectively yield, area under wheat and area under cultivation see Singh et al. (2004).

Assuming that the population mean \( \bar{Z} \) of the variable \( z \) is known, Chand (1975) proposed a chain type ratio estimator as

\[
t_1 = \frac{\bar{y}}{\bar{x} \left( \frac{\bar{x}}{\bar{Z}} \right)} \bar{Z} \tag{1.3}
\]

Several authors have used prior value of certain population parameter(s) to find more precise estimates. Singh and Upadhyaya (1995) used coefficient of variation of \( z \) for defining
modified chain type ratio estimator. In many situations the value of the auxiliary variable may be available for each unit in the population, for instance, see Das and Tripathi (1981). In such situations knowledge on $\bar{Z}$, $C_z$, $\beta_1(z)$ (coefficient of skewness), $\beta_2(z)$ (coefficient of kurtosis) and possibly on some other parameters may be utilized. Regarding the availability of information on $C_z$, $\beta_1(z)$ and $\beta_2(z)$, the researchers may be referred to Searls(1964), Sen(1978), Singh et al.(1973), Searls and Intarapanich(1990) and Singh et al.(2007). Using the known coefficient of variation $C_z$ and known coefficient of kurtosis $\beta_2(z)$ of the second auxiliary character $z$ Upadhyaya and Singh (2001) proposed some estimators for $\bar{Y}$.

If the population mean and coefficient of variation of the second auxiliary character is known, the standard deviation $\sigma_z$ is automatically known and it is more meaningful to use the $\sigma_z$ in addition to $C_z$, see Srivastava and Jhajj (1980). Further, $C_z$, $\beta_1(z)$ and $\beta_2(z)$ are the unit free constants, their use in additive form is not much justified. Motivated with the above justifications and utilizing the known values of $\sigma_z$, $\beta_1(z)$ and $\beta_2(z)$, Singh (2001) suggested some modified estimators for $\bar{Y}$.

In this paper, under simple random sampling without replacement (SRSWOR), we have suggested improved chain ratio type estimator for estimating population mean using some known values of population parameter(s).

2. The suggested estimator

The work of authors discussed in section 1 can be summarized by using following estimator

\[
t = \bar{y} \left( \frac{\bar{x}'}{\bar{x}} \right) \left( \frac{a\bar{Z} + b}{a\bar{Z}' + b} \right)
\]

where $a(\neq 0)$, $b$ are either real numbers or the functions of the known parameters of the second auxiliary variable $z$ such as standard deviation ($\sigma_z$), coefficient of variation ($C_z$), skewness ($\beta_1(z)$) and kurtosis ($\beta_2(z)$).

The following scheme presents some of the important known estimators of the population mean which can be obtained by suitable choice of constants $a$ and $b$. 
In addition to these estimators a large number of estimators can also be generated from the estimator $t$ at (2.1) by putting suitable values of $a$ and $b$.

Following Kadilar and Cingi (2006), we propose modified estimator combining $t_1$ and $t_i$ ($i = 2,3,...,7$) as follows
\[ t_i^* = \alpha t_i + (1 - \alpha) t_i, \quad (i = 2, 3, \ldots, 7) \]  

(2.2)

where \( \alpha \) is a real constant to be determined such that MSE of \( t_i^* \) is minimum and \( t_i \)  
\( (i = 2, 3, \ldots, 7) \) are estimators listed above.

To obtain the bias and MSE of \( t_i^* \), we write

\[
\begin{align*}
\bar{y} &= \overline{Y}(1 + e_0), \quad \bar{x} = \overline{X}(1 + e_1), \quad \bar{x}' = \overline{X}(1 + e'_1), \quad \bar{z}' = \overline{Z}(1 + e'_2)
\end{align*}
\]

such that

\[
E(e_0) = E(e_1) = E(e'_1) = E(e'_2) = 0
\]

and

\[
E(e_0^2) = f_1 C_y^2, \quad E(e_1^2) = f_1 C_x^2, \quad E(e'_1^2) = f_2 C_x^2
\]

\[
E(e_2^2) = f_2 C_z^2, \quad E(e_0 e_1) = f_1 \rho_{xy} C_x C_y, \quad E(e_0 e'_1) = f_2 \rho_{xy} C_x C_y
\]

\[
E(e_0 e'_2) = f_2 \rho_{yz} C_y C_z, \quad E(e_1 e'_2) = f_2 \rho_{xz} C_x C_z
\]

\[
E(e'_1 e'_2) = f_2 \rho_{xz} C_x C_z
\]

where

\[
f_1 = \left( \frac{1}{n} - \frac{1}{N} \right), \quad f_2 = \left( \frac{1}{n'} - \frac{1}{N} \right),
\]

\[
C_y^2 = \frac{S_y^2}{\overline{Y}^2}, \quad C_x^2 = \frac{S_x^2}{\overline{X}^2}, \quad C_z^2 = \frac{S_z^2}{\overline{Z}^2}
\]

\[
\rho_{xy} = \frac{S_{xy}}{S_x S_y}, \quad \rho_{xz} = \frac{S_{xz}}{S_x S_z}, \quad \rho_{yz} = \frac{S_{yz}}{S_y S_z}
\]

\[
S_y^2 = \frac{1}{(N-1)} \sum_{i \in U} (y_i - \overline{Y})^2, \quad S_x^2 = \frac{1}{(N-1)} \sum_{i \in U} (x_i - \overline{X})^2
\]

\[
S_z^2 = \frac{1}{(N-1)} \sum_{i \in U} (z_i - \overline{Z})^2, \quad S_{xy} = \frac{1}{(N-1)} \sum_{i \in U} (x_i - \overline{X})(y_i - \overline{Y})
\]
Expressing $t^*$ in terms of e's, we have
\[
t^*_i = \bar{Y}(1 + e_0)(\alpha(1 + e'_i)(1 + e_1)^{-1} + (1 - \alpha)(1 + e'_i)(1 + e_1)^{-1}(1 + \theta e'_i)^{-1})
\] (2.3)

where $\theta = \frac{aZ}{aZ + b}$ (2.4)

Expanding the right hand side of (2.3) and retaining terms up to second power of e's, we have
\[
t^*_i \simeq \bar{Y}[1 + e_0 - e_1 + e'_i - e'_i(\alpha + \theta - \alpha \theta)]
\] (2.5)
or
\[
t^*_i - \bar{Y} \simeq \bar{Y}[e_0 - e_1 + e'_i - e'_i(\alpha + \theta - \alpha \theta)]
\] (2.6)

Squaring both sides of (2.6) and then taking expectation, we get the MSE of the estimator $t^*_i$, up to the first order of approximation, as
\[
\text{MSE}(t^*_i) = \bar{Y}^2\left[f_3 C_y^2 + f_3 C_x^2 + (\alpha + \theta - \alpha \theta)^2 f_2 C_z^2 - 2f_3 \rho C_y C_x - 2(\alpha + \theta - \alpha \theta)f_3 \rho C_y C_z \right]
\] (2.7)

where
\[
f_3 = \left(\frac{1}{n} - \frac{1}{n^2}\right).
\]

Minimization of (2.7) with respect to $\alpha$ yield its optimum value as
\[
\alpha_{opt} = \frac{K_{yz} - \theta}{1 - \theta}
\] (2.8)

where
\[
K_{yz} = \rho_{yz} \frac{C_y}{C_z}.
\]
Substitution of (2.8) in (2.7) yields the minimum value of MSE \( (\tau^*_i) \) as

\[
\min \text{MSE}(\tau^*_i) = M_a = \bar{Y}^2 \left[ f_y^2 C_x^2 + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) - f_y^2 \rho_{yx}^2 C_y^2 \right] 
\]  

(2.9)

3. Efficiency comparisons

In this section, the conditions for which the proposed estimator is better than \( \tau_i (i = 1, 2, ..., 7) \) have been obtained. The MSE’s of these estimators up to the order \( o(n)^{-1} \) are derived as

\[
\text{MSE}(\bar{Y}_{rd}) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.1)

\[
\text{MSE}(\tau_1) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.2)

\[
\text{MSE}(\tau_2) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_2 C_x^2 - 2 \theta_2 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.3)

\[
\text{MSE}(\tau_3) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_3 C_x^2 - 2 \theta_3 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.4)

\[
\text{MSE}(\tau_4) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_4 C_x^2 - 2 \theta_4 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.5)

\[
\text{MSE}(\tau_5) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_5 C_x^2 - 2 \theta_5 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.6)

\[
\text{MSE}(\tau_6) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_6 C_x^2 - 2 \theta_6 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.7)

\[
\text{MSE}(\tau_7) = \bar{Y}^2 \left[ f_y C_y^2 + f_y (\theta_7 C_x^2 - 2 \theta_7 \rho_{yx} C_y C_x) + f_y (C_x^2 - 2 \rho_{yx} C_y C_x) \right] 
\]  

(3.8)

and

\[
\text{MSE}(\bar{Y}_{rd}) - M_o = f_y \rho_{yx}^2 C_y^2 \geq 0 
\]  

(3.9)

From (2.9) and (3.1), we have

\[
\text{MSE}(\bar{Y}_{rd}) - M_o = f_y \rho_{yx}^2 C_y^2 \geq 0 
\]  

(3.9)

Also from (2.9) and (3.2)-(3.8), we have
Thus it follows from (3.9) and (3.10) that the suggested estimator under optimum condition is always better than the estimator \( t_i, (i = 1, 2, \ldots, 7) \).

4. Empirical study

To illustrate the performance of various estimators of \( \overline{Y} \), we consider the data used by Anderson (1958). The variates are

\begin{align*}
\begin{array}{ll}
y & : \text{Head length of second son} \\
x & : \text{Head length of first son} \\
z & : \text{Head breadth of first son}
\end{array}
\end{align*}

\[
N = 25, \quad \overline{Y} = 183.84 =, \quad \overline{X} = 185.72, \quad \overline{Z} = 151.12, \quad \sigma_z = 7.224, \quad C_y = 0.0546, \\
C_x = 0.0526, \quad C_z = 0.0488, \quad \rho_{yx} = 0.7108, \quad \rho_{yz} = 0.6932, \quad \rho_{xz} = 0.7346, \quad \beta_1(z) = 0.002, \quad \beta_2(z) = 2.6519.
\]

Consider \( n' = 10 \) and \( n = 7 \).

We have computed the percent relative efficiency (PRE) of different estimators of \( \overline{Y} \) with respect to usual estimator \( \overline{Y} \) and compiled in the table 4.1:

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
estimator & PRE \\
\hline
\( \overline{Y} \) & 100 \\
\hline
\( \overline{Y}_{rd} \) & 122.5393 \\
\hline
\( t_1 \) & 178.8189 \\
\hline
\( t_2 \) & 178.8405 \\
\hline
\( t_3 \) & 178.8277 \\
\hline
\( t_4 \) & 186.3912 \\
\hline
\( t_5 \) & 181.6025 \\
\hline
\( t_6 \) & 122.5473 \\
\hline
\end{tabular}
\caption{PRE of different estimators of \( \overline{Y} \) with respect to \( \overline{Y} \)}
\end{table}
5. Conclusion

We have suggested modified estimators $t_i^*(i = 2,3,\ldots,7)$. From table 4.1, we conclude that the proposed estimators are better than usual two-phase ratio estimator $\bar{y}_{rd}$, Chand (1975) chain type ratio estimator $t_1$, estimator $t_2$ proposed by Singh and Upadhyaya (1995), estimators $t_i(i = 3,4)$ and than that of Singh (2001) estimators $t_i(i = 5,6,7)$. For practical purposes the choice of the estimator depends upon the availability of the population parameter(s).

References


